ISTITUTO ITALIANO
DI TECNOLOGIA

# All You Should Know on Visual Recognition Pipelines: from theory to iCub

**Sean Ryan Fanello**
iCub Facility – Istituto Italiano di Tecnologia

# Motivations


**Self-Supervised Strategies**

**Kinematics**

**Motion**

Human Robot Interaction is a **new** and **natural** application for object recognition

In robotics settings strong cues are often available, therefore object detectors can be avoided

Recognition as tool for complex tasks: grasp, manipulation, affordances, pose

S.R. Fanello, C. Ciliberto, L. Natale, G. Metta – Weakly Supervised Strategies for Natural Object Recognition in Robotics, ICRA 2013

C. Ciliberto, S.R. Fanello, M. Santoro, L. Natale, G. Metta, L. Rosasco - On the Impact of Learning Hierarchical Representations for Visual Recognition in Robotics, IROS 2013

# Introduction

## Single Instance Recognition

Image

Recognition System

Pattern Recognition & Machine Learning By C. Bishop

## Object Categorization

Image

Recognition System

A book

## Image Retrieval

Image

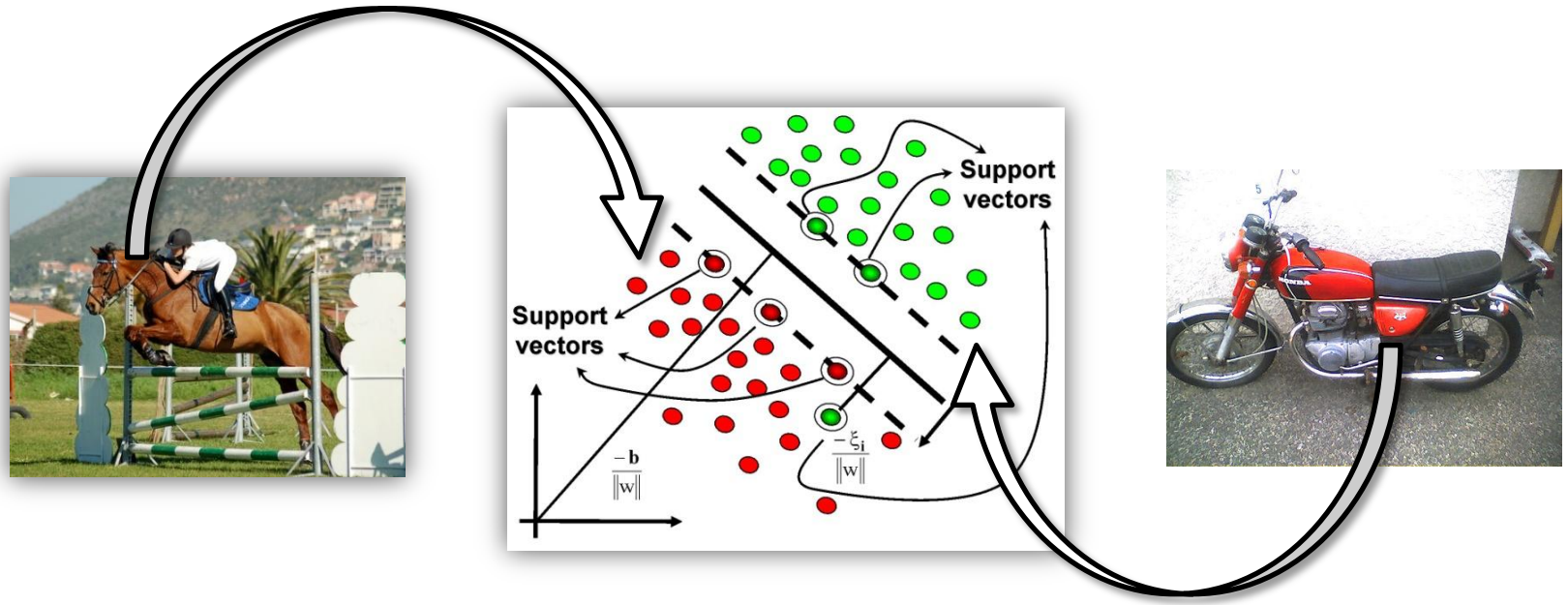
Recognition System


"Same" Image Representations

# Introduction

- Geometric Information
- *Invariance to Image Transformation*
- *Discriminative Power*
- *Real-Time*

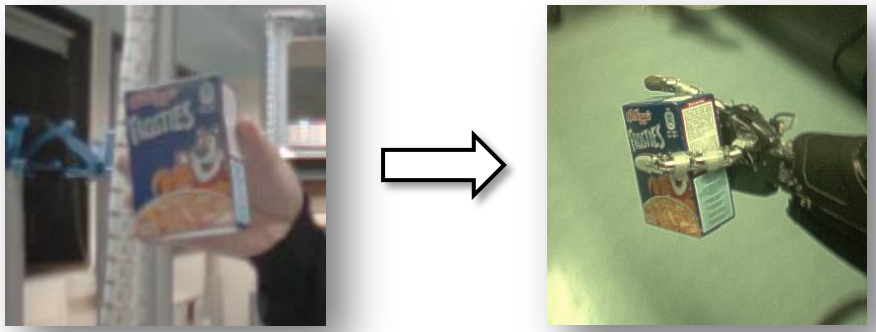
***Still an open problem…***



# Image Description

We want to convert an image into a compact feature vector $\mathbf{v} \in \mathbb{R}^n$

It must be robust to viewpoint, illumination, occlusion etc.
Ideally an Invariant Representation



Invariant to: scale, occlusions, lighting, view-point

Example: Color Histogram

Nice invariance properties, but not informative
Trade-off between invariance and information

**More Data ⟹ Invariance is not needed**

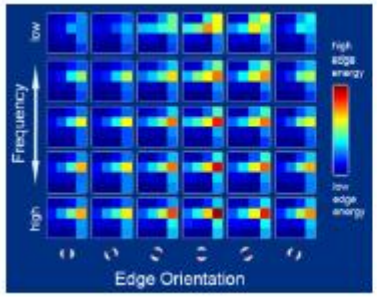
# Global Descriptors

## Color Histogram

Swain, Ballard, "Color indexing", IJCV'91.

## GIST of a Scene

Oliva, Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope", IJCV'01.
Douze, Jegou, Sandhawalia, Amsaleg, Schmid, "Evaluation of GIST descriptors for web-scale image search", CIVR'09.




## CENTRIST: Census Transform hISTogram

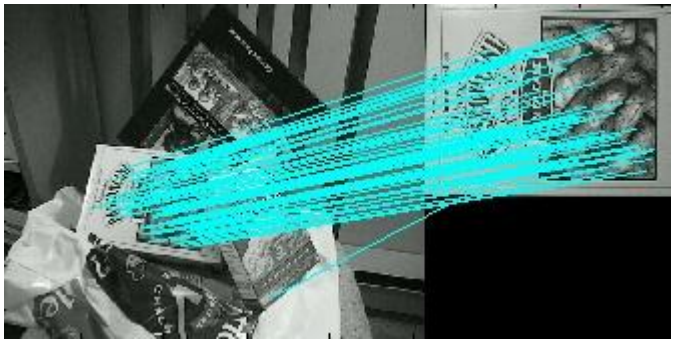
Wu, Rehg, "CENTRIST: a visual descriptor for scene categorization", TPAMI'11.

**Highly efficient to compute and to match, but they lack of description power. Not suitable for discriminative tasks**

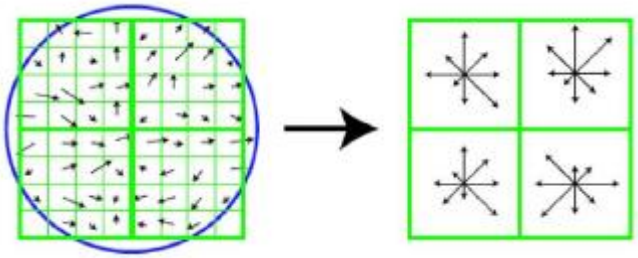
# Local Descriptors

A set of local descriptors is extracted. These features are invariant to geometric transformations. Most widely used: SIFTs

Lowe, "Distinctive image features from scale-invariant keypoints", IJCV'04.

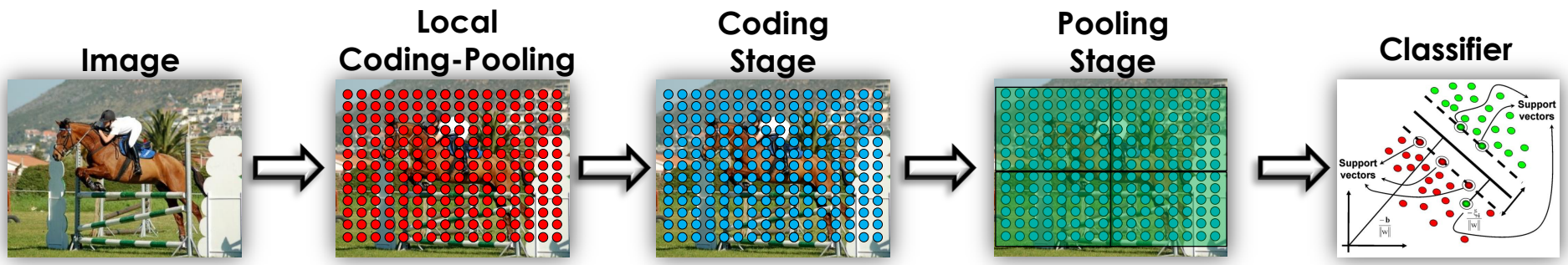


8 orientations of the gradient in 4x4 spatial grid. 128 dimensions



**Notice: Local features are invariant when both keypoint detector and descriptor are used. In object categorization detectors do not perform well due to intra-class variability**

# Visual Recognition Pipelines



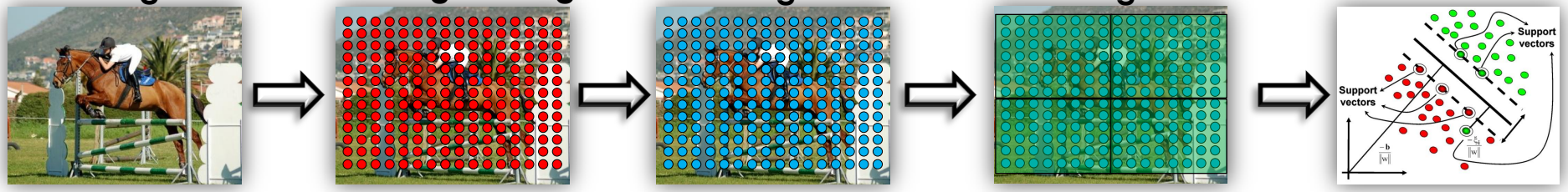

Current state-the-art recognition pipelines use hierarchical image representations.

Main Idea:
Starting from low-level descriptors (e.g. SIFT) we compute higher order statistics with a sequence of coding-pooling operators.

BOW-like systems have the first layer hand crafted (e.g. SIFT)
Pro: Efficient, easy to implement Cons: No invariance

Deep Architectures (e.g. HMAX) try to learn also the first layer
Pro: Invariant to small transformation, Higher (?) accuracy Cons: TOO slow

# Visual Recognition Pipelines




i-th Object Class

e.g. SIFT, HOG, Patch

e.g. VQ, SC, LLC, VLAD, SV, FK

e.g. SPR

e.g. SVM. MRF, BN

***Local Descriptors:***

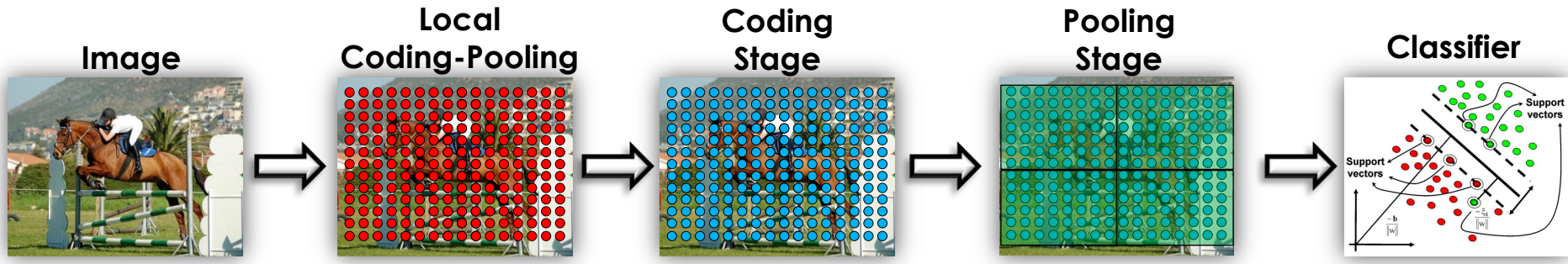
$\mathbf{x}_1, \ldots, \mathbf{x}_M$ are extracted from the image

Keypoint detectors are not suitable, a dense grid better catches image statistics.

These descriptors are not invariant to scale, rotation and translation etc.

Main Assumption:
Enough Data is available

# Visual Recognition Pipelines



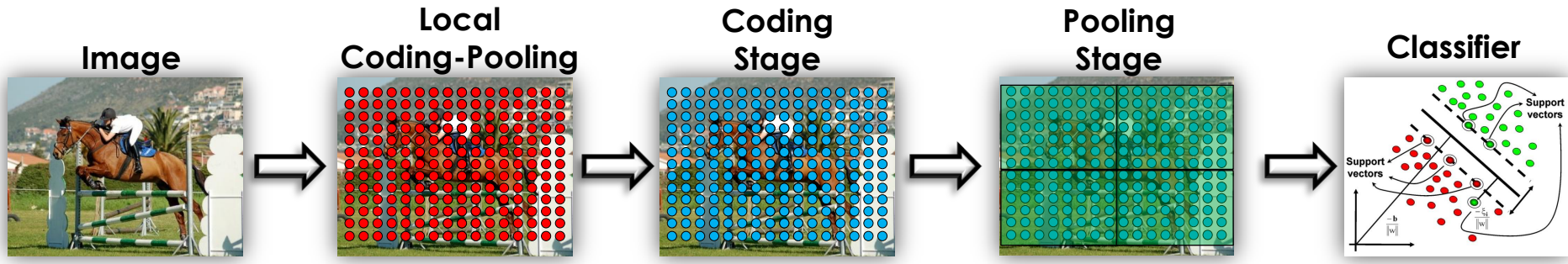

***Dictionary Learning*:**

We want to relate local descriptors to a common basis, called dictionary. Intuitively a dictionary $\mathbf{D} = [\mu_1, \ldots, \mu_K]$ represents all the relevant descriptors and it is able to describe the content of any image. The dictionary is learned using K-Means algorithm with K=dictionary size on a large set of SIFTs (~1M)

$$\min_{\mathbf{D},\mathbf{U}} \|\mathbf{X} - \mathbf{D}\mathbf{U}\|_F^2$$

$$\text{s.t.} \quad \text{Card}(\mathbf{u}_i) = 1, |\mathbf{u}_i| = 1, \mathbf{u}_i \succeq 0, \forall i = 1, \ldots, T$$

# Visual Recognition Pipelines




## *Coding Operators*

The coding stage maps the input features $\mathbf{x}_1, \ldots, \mathbf{x}_M$ into an overcomplete space.

### *Based on the Reconstruction Error*

$$\mathbf{u}_i = \arg\min_{\mathbf{u}} \|\mathbf{x} - \mathbf{D}\mathbf{u}\|^2 + \lambda \mathrm{R}(\mathbf{u})$$
$$\text{s.t.} \quad \mathrm{C}(\mathbf{u})$$

### *Higher Order Statistics*

$$\mathbf{u_k} = \frac{1}{M\sqrt{\pi_k}} \sum_{i=1}^{M} q_{ik} \Sigma_k^{-\frac{1}{2}} (\mathbf{x}_i - \mu_k)$$

$$\mathbf{v_k} = \frac{1}{M\sqrt{2\pi_k}} \sum_{i=1}^{M} q_{ik} [(\mathbf{x}_i - \mu_k)^{\mathrm{T}} \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) - 1]$$

# Coding Operators

**Vector Quantization (VQ)**

$$\min_{\mathbf{u}_i} \|\mathbf{x}_i - \mathbf{D}\mathbf{u}_i\|^2$$
$$\text{s.t.} \quad \text{Card}(\mathbf{u}_i) = 1, |\mathbf{u}_i| = 1, \mathbf{u}_i \succeq 0$$

**Sparse Coding (SC)**

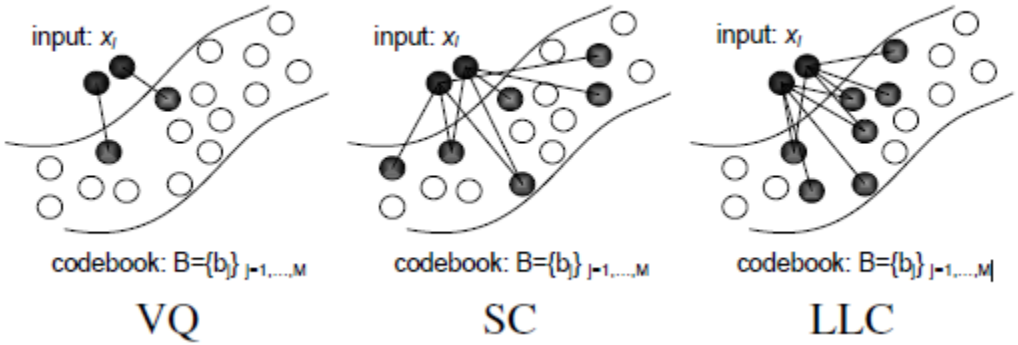
$$\min_{\mathbf{u}_i} \|\mathbf{x}_i - \mathbf{D}\mathbf{u}_i\|^2 + \lambda\|\mathbf{u}_i\|_1$$

**Locality-constrained Linear Coding (LLC)**

$$\min_{\overline{\mathbf{u}}_i} \|\mathbf{x}_i - \overline{\mathbf{D}}\overline{\mathbf{u}}_i\|^2$$
$$\text{s.t.} \quad 1^{\mathrm{T}}\overline{\mathbf{u}}_i = 1$$

**They all minimize the reconstruction error:**

$$\mathbf{u}_i = \arg\min_{\mathbf{u}} \|\mathbf{x} - \mathbf{D}\mathbf{u}\|^2 + \lambda \mathrm{R}(\mathbf{u})$$
$$\text{s.t.} \quad \mathrm{C}(\mathbf{u})$$

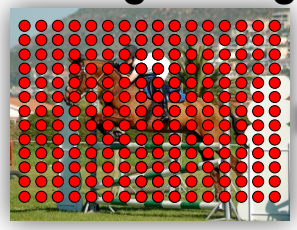


# Visual Recognition Pipelines

Image

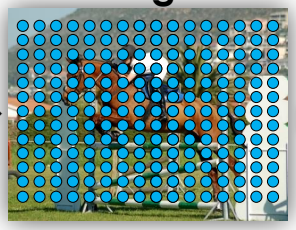
i-th Object Class

Local Coding-Pooling



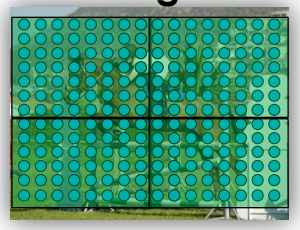
e.g. SIFT, HOG, Patch

Coding Stage



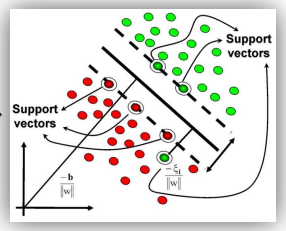
e.g. VQ, SC, LLC, VLAD, SV, FK

Pooling Stage



e.g. SPR

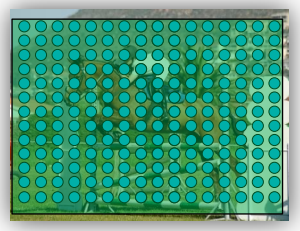
Classifier



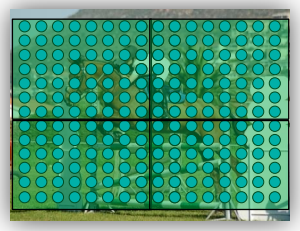

e.g. SVM. MRF, BN

***Spatial Pyramid Representation***

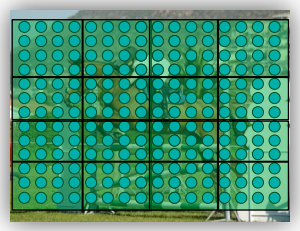
*Image partitioned in $2^l \times 2^l$ segments*



l=0



l=1



l=2

21 Spatial Bins

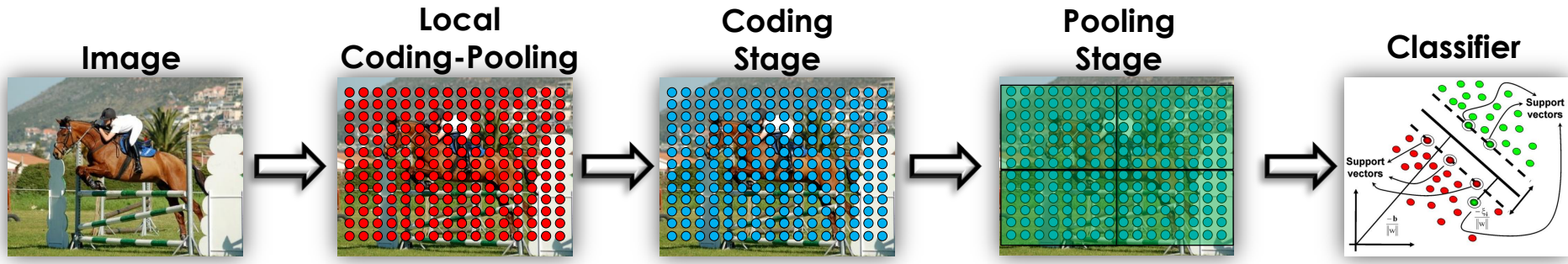
***Pooling Operators***

They aggregates local descriptors into a single one.

***Example: Max Pooling***

$$\mathbf{h}_{s,j} = \max_{i \in Y_s} u_{i,j} \quad \forall j = 1, \ldots, K$$

# Visual Recognition Pipelines




***Learning Stage***

After the pooling stage we get a descriptor $\mathbf{z} \in \mathbb{R}^{KS}$ where K is the dictionary size and S the number of cells of the spatial pyramid

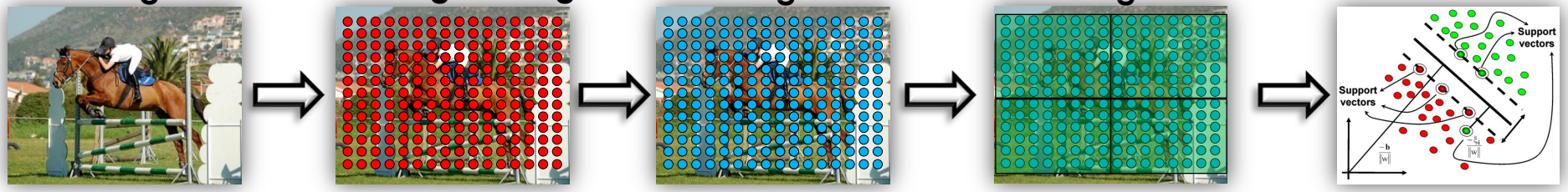
Multi-class classification problem that can be solved with a standard One-vs-All paradigm. Linear Classifiers are more suitable for Real-Time tasks

$$\underset{\mathbf{W} \in \mathbb{R}^{d \times k}}{\text{Minimize}} \quad \lambda\, \Omega(\mathbf{W}) + \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, \mathbf{W}^T \mathbf{x}_i\right)$$

Non-linear feature mapping can be used to approximate non-linear Kernels

A. Vedaldi, A. Zisserman – Efficient additive kernels via explicit feature maps. CVPR 2010

# Recap




***Local Descriptors***

Dense Grid of SIFT descriptors with fixed scale and grid spacing

***Dictionary Learning***

A dictionary is learned with K-Means (K=dictionary size) on a large set of SIFTs

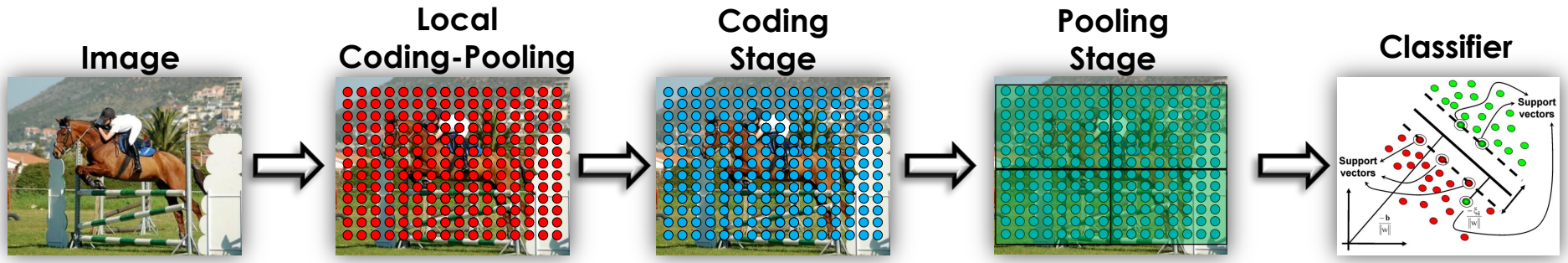
***Coding-Pooling***

Features are mapped into an overcomplete space and then aggregated together. The image is transformed into a single vector $\mathbf{v} \in \mathbb{R}^n$

***Linear Classification***

Use your favorite classification method (GURLS!) with One-vs-All paradigm

# Generalizing the Pipeline




i-th Object Class

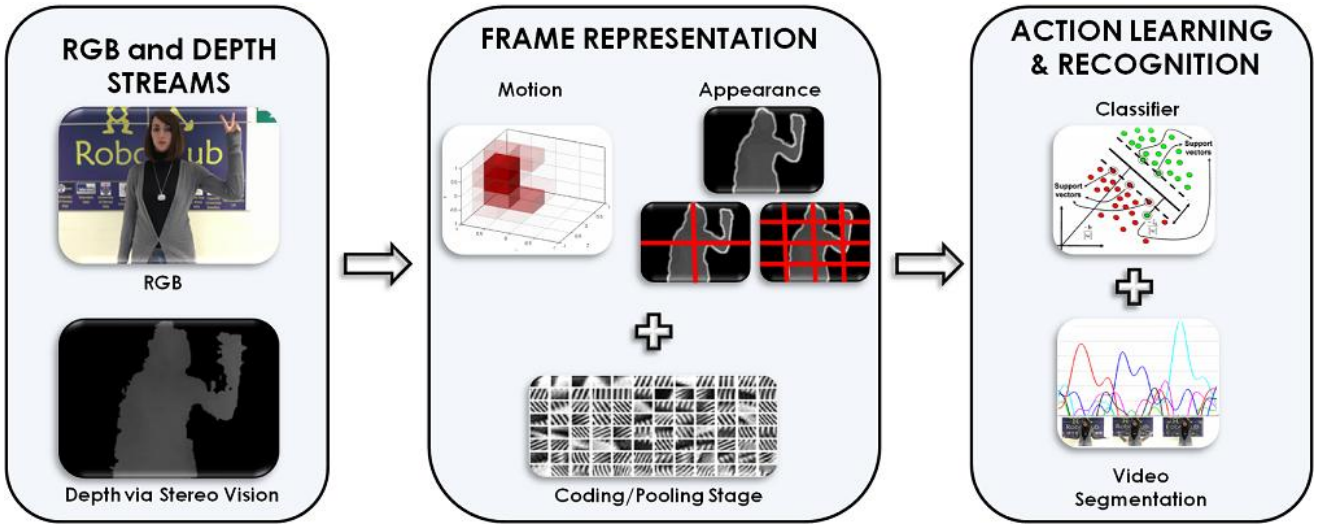
e.g. SIFT, HOG, Patch

e.g. VQ, SC, LLC, VLAD, SV, FK

e.g. SPR

e.g. SVM. RLS, MRF, BN

***Action/Gesture Recognition***

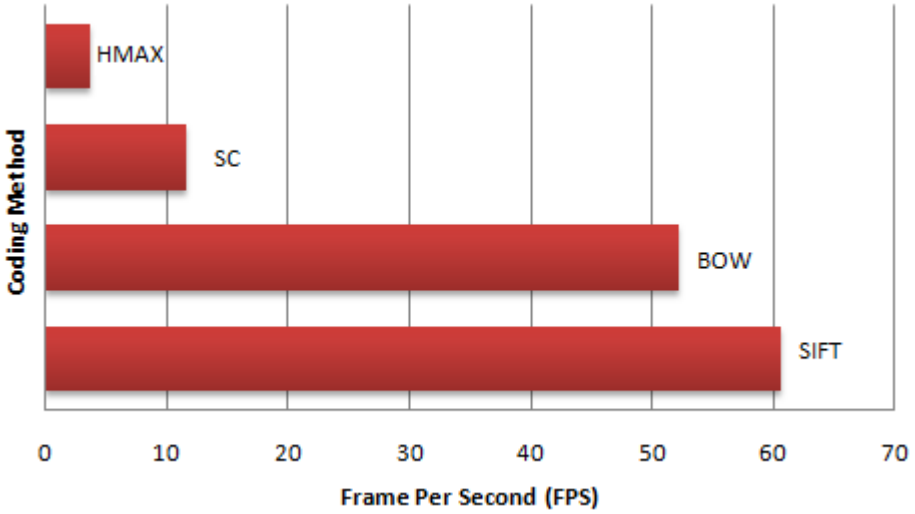


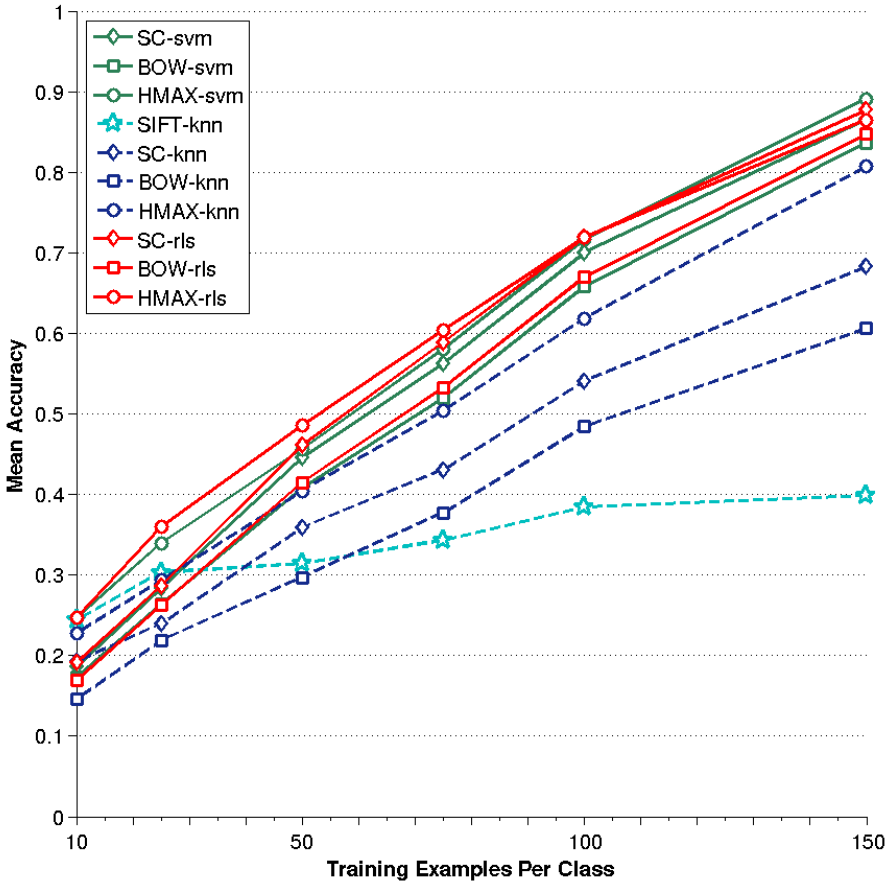

## *Main Modifications*

- *Image Representation from stereo depth*
- *Motion and Appearance*
- *Temporal Pooling*
- *Video Segmentation*

# In The Real World...

## iCubWorld Single Instance Recognition




C. Ciliberto, S.R. Fanello, M. Santoro, L. Natale, G. Metta, L. Rosasco - On the Impact of Learning Hierarchical Representations for Visual Recognition in Robotics, IROS 2013

# iCubWorld Data-Set

**Self-Supervised Strategies**

**Kinematics**

**Motion**

**iCubWorld Categorization Data-Set**
**10 Categories, 40 Objects**


Object Categorization between Computer Vision & Robotics
**Main Difficulty: structured clutter**, therefore the context cannot be exploited.

**iCubWorld available at: http://www.iit.it/it/projects/data-sets.html**

S.R. Fanello, C. Ciliberto, M. Santoro, L. Natale, G. Metta, L. Rosasco, F. Odone – iCubWorld: Friendly Robots Help Building Good Vision Data-Sets, CVPRW 2013

# Build your own app


***Example of application with visual recognition capabilities***

# SparseCoder Module

***Where?***
$ICUB_ROOT/contrib/src/sparseCoder

***Dependencies***
SIFT GPU, OpenCV 2.X, Yarp

***Docs:***
http://wiki.icub.org/iCub/contrib/dox/html/group__icub__image__representation.html

***Coding Methods:***
Best Code Entries, Sparse Coding, Bag of Words

***Spatial Pyramid Matching:***
Yes, customizable pyramid levels

***Dictionary Learning:***
With K-Means

***More Functionalities:***
PCA on SIFTs, Power Normalization, Sparse & Dense SIFTs

# Thank You!
# Any Question?